Monaural separation of dependent audio sources based on a generalized Wiener filter

Ma, Guilin; Agerkvist, Finn T.; Luther, J.B.

Published in:
Proceedings of the 7th IEEE International Symposium on Signal Processing and Information Technology

Link to article, DOI:
10.1109/ISSPIT.2007.4458074

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

2007 IEEE International Symposium
on Signal Processing and Information Technology

Monaural Separation of Dependent Audio Sources Based on a Generalized Wiener Filter

Guilin Ma[1], Finn T. Agerkvist[1], Jim Benjamin Luther[2]
[1] Acoustic Technology, Ørsted, Technical University of Denmark, Building 352, 2800 Lyngby, Denmark.
[2] GN ReSound Denmark, Lautrupbjerg 9, 2750 Ballerup, Denmark.

Abstract *– This paper presents a two-stage approach for single-channel separation of dependent audio sources. The proposed algorithm is developed in the Bayesian framework and designed for general audio signals. In the first stage of the algorithm, the joint distribution of discrete Fourier transform (DFT) coefficients of the dependent sources is modeled by complex Gaussian mixture models in the frequency domain from samples of individual sources to capture the properties of the sources and their correlation. During the second stage, the mixture is separated through a generalized Wiener filter, which takes correlation term and local stationarity into account. The performance of the algorithm is tested on real audio signals. The results show that the proposed algorithm works very well when the dependent sources have comparable variances and linear correlation.*

Keywords – *monaural source separation, complex Gaussian mixture model, Gaussian statistical model of DFT coefficients*

I. INTRODUCTION

Source separation problem arises in a variety of signal processing applications. It can be categorized in several ways: Depending on the amount of available information about the mixing process and sources, it can be divided into blind source separation (BSS) and semi-BSS; According to the relation of n (the number of sources) and m (the number of sensors), it falls into the categories of an under-determined problem ($m < n$), even-determined problem ($m = n$) and over-determined problem ($m > n$); Based on the relation between sources, it is either a problem with independent sources or a problem with dependent sources.

Most of the source separation algorithms are based on the assumption that the sources are statistically independent, which holds in most cases. However, in some special audio applications such as feedback cancellation in hearing aids, the mixture contains dependent sources. The very few algorithms dealing with dependent sources include both semi-BSS techniques and BSS techniques. The semi-BSS techniques are heavily dependent on the nature of the problem and thus very ad-hoc. A typical example is presented in [1], where the structure of the mixing matrix and source covariance matrices are known beforehand. Instead of applying strong prior knowledge, the BSS techniques usually make strong assumptions on the properties of the sources [2][3][4], such as time-frequency sparsity, to solve the dependent source separation problem. When the sources are only linearly correlated through room impulse responses, the problem reduces to a convolutive BSS problem, which is a very active and challenging research area.

Compared with even- and over-determined problems, under-determined source separation problems are generally much more difficult due to the lack of constraints. Additional constraints are normally applied by making strong assumptions on the source characteristics, incorporating sources models or providing prior knowledge on the mixing process and/or signals. One powerful assumption about the sources is that they have a parsimonious representation in a given basis, such as the time-frequency (T-F) representation. This kind of assumption has lead to encouraging techniques [4][5][6]. Another class of methods incorporate source models, such as Vector Quantization (VQ), Gaussian Mixture Models (GMM), train the models first and separate the mixture afterwards based on proper criteria (e.g., minimum mean square error, likelihood ratio) [7][8].

The problem discussed in this paper, monaural separation of two general audio signals that are strongly dependent, is a combination of an under-determined problem and a problem with correlated sources. Most of the algorithms reviewed above fail in this extreme case either because the sparse representation for correlated sources is not valid or because the algorithms require multiple channels. For example, the existing technique for under-determined convolutive BSS requires multiple channels [9].

Since the sources are general, specific source models, such as speech model, are not applicable. Besides, how the two signals are correlated and what properties the sources exhibit are unknown. To combat these difficulties, a two-stage algorithm based on generalized Wiener filtering is proposed in this paper. In the first step, complex GMM is exploited to acquire sufficient knowledge about the sources and the way they are correlated. Based on the information obtained in the first step, the mixture is separated later by a generalized Wiener filter.

Although the study here is for two dependent sources, the method proposed can be generalized to more sources at least theoretically.

II. PROBLEM FORMULATION

The microphone signal x is a mixture of two dependent signals s_1 and s_2, i.e.,

$$x = s_1 + s_2 \tag{1}$$

In the Bayesian framework, the two sources can be estimated through estimators such as maximum likelihood (ML). However, since the problem is under-determined, there will be multiple solutions with the ML estimator [7]. One of the alternatives is the maximum *a posteriori* (MAP) estimator:

$$\begin{aligned}(\hat{s}_1, \hat{s}_2) &= \arg\max_{s_1, s_2} p(s_1, s_2 \,|\, x) \\ p(s_1, s_2 \,|\, x) &\propto p(x \,|\, s_1, s_2)\, p(s_1, s_2)\end{aligned} \tag{2}$$

where $p(x|s_1, s_2)$ is the likelihood function, $p(s_1, s_2)$ is the prior knowledge about the joint distribution of the sources, which essentially reflects the statistical properties of each individual source and the correlation of the two sources.

A similar estimator is the conditional posterior mean (PM):

$$(\hat{s}_1, \hat{s}_2) = E\left[s_1, s_2 \,|\, x\right] \tag{3}$$

where the expectation operator $E[\bullet]$ implicitly requires the knowledge of the joint distribution $p(s_1, s_2)$.

Therefore, in the Bayesian framework, the solution for this source separation problem includes a stage of estimating the joint distribution and a second stage of separating the mixture based on a proper estimator such as MAP or PM.

III. ESTIMATION OF JOINT DISTRIBUTION

The estimation of the joint distribution can be performed in the time domain or any domain spanned by proper basis functions. Since the correlation between the two dependent sources usually varies strongly with frequency, time-domain modeling lacks the resolution to describe the difference among frequency bins and consequently leads to degraded performance. For discrete signals, discrete Fourier basis has several desirable properties and serves as an efficient domain for analyzing the signals in this paper.

A. Gaussian Statistical Model of DFT Coefficients

For real-time processing purpose, the sampled microphone signal $x(n)$ is broken into frames. Each frame is Fourier transformed. This process is referred to as short-time Fourier transform (STFT), i.e.,

$$\begin{aligned}\tilde{X}_k(m) &= \sum_{n=0}^{L-1} x((m-1)(L-M)+n+1)h(n)e^{-j\frac{2\pi}{L}kn} \\ k &= 0, 1, \cdots, L-1 \qquad m = 1, 2, \cdots\end{aligned} \tag{4}$$

where L is the length of each frame, M is the length of overlapping, m is the frame index, $h(n)$ is the window function applied. $\tilde{X}_0(m)$ and $\tilde{X}_{L/2}(m)$ are not interesting since they are direct current (DC) and Nyquist components respectively. $\tilde{X}_k(m)\ \ k = L/2+1, \cdots, L-1$ are also ignored due to the symmetry of DFT coefficients. A tilde is used to denote a complex quantity. For brevity, m is dropped out in the following formulas, and $h(n)$ is also neglected.

A widely accepted assumption for stationary audio signals is that the DFT coefficients are statistically independent Gaussian random variables [11], i.e.,

$$\begin{aligned}&\begin{bmatrix}\tilde{X}_1 \\ \vdots \\ \tilde{X}_{\frac{L}{2}-1}\end{bmatrix} \sim CN(\boldsymbol{0}, diag\{\sigma_1^2, \cdots, \sigma_{\frac{L}{2}-1}^2\}) \\ &\sigma_k^2 = Var(\tilde{X}_{k,re}) + Var(\tilde{X}_{k,im})\end{aligned} \tag{5}$$

where symbols with bold and italic font represent matrices or vectors, $diag\{\bullet\}$ is the diagonal matrix formed by the listed entries, $Var(\bullet)$ is the variance of the listed entries, subscripts 're' and 'im' denote the real part and imaginary part of a complex quantity respectively, and *CN* denotes the complex Gaussian distribution [10].

(5) implies that the DFT coefficients in different frequency bins are independent. It also implies that the real and imaginary parts of coefficients in each frequency bin are independent, have Gaussian distributions and the same variances. In a strict sense, the DFT coefficients follow an asymptotical Gaussian distribution as L approaches infinity [11].

To reduce the large number of parameters to estimate, it is assumed that the two correlated zero-mean stationary signals $s_1(n)$ and $s_2(n)$ are only correlated within the same frequency bin as shown in (6). This assumption usually holds very well for many types of correlation, especially linear correlation.

$$\begin{aligned}&\tilde{S}_{1,k_1,2,k_2} \sim CN(\boldsymbol{0}, \begin{bmatrix}\sigma_{1,k_1}^2 & 0 \\ 0 & \sigma_{2,k_2}^2\end{bmatrix}) \qquad k_1 \neq k_2 \\ &\tilde{S}_{1,k,2,k} \sim CN(\boldsymbol{0}, \begin{bmatrix}\sigma_{1,k}^2 & \sigma_{12,k} \\ \sigma_{12,k}^* & \sigma_{2,k}^2\end{bmatrix}) \qquad \sigma_{12,k} = E[\tilde{S}_{1,k}^* \tilde{S}_{2,k}] \\ &\tilde{S}_{1,k_1,2,k_2} = [\tilde{S}_{1,k_1}, \tilde{S}_{2,k_2}]^T \\ &k, k_1, k_2 = 1, \cdots, L/2-1\end{aligned} \tag{6}$$

To obtain an expressible probability density function in terms of $\tilde{S}_{1,k_1,2,k_2}$, we have to further assume a special relation between the covariance matrices of $\tilde{S}_{1,k_1}$ and $\tilde{S}_{2,k_2}$ [12]:

$$\begin{aligned}&Cov(\tilde{S}_{1,k_1,re}, \tilde{S}_{2,k_2,re}) = Cov(\tilde{S}_{1,k_1,im}, \tilde{S}_{2,k_2,im}) \\ &Cov(\tilde{S}_{1,k_1,re}, \tilde{S}_{2,k_2,im}) = -Cov(\tilde{S}_{2,k_2,re}, \tilde{S}_{1,k_1,im}) \\ &k_1, k_2 = 1, \cdots, L/2-1\end{aligned} \tag{7}$$

where $Cov(\bullet)$ is the covariance of the listed entries. It was found that the assumption (7) holds well at least for linear correlation.

B. GMM Estimation of the Joint Distribution

As seen above, the DFT coefficients of a stationary audio signal can be described by Gaussian distributions. Therefore, its power spectral density, which gives the variance as a

function of frequency, is completely taken into account by the Gaussian distributions. However, realistic audio signals are only locally stationary and contain various types of timbres and pitches [13]. The complex GMM, instead of a single complex Gaussian distribution, has to be adopted to capture the diverse spectra of the signals. As a semi-parametric method to estimate the probability density function, GMM also possesses the advantages of high flexibility and reasonable complexity compared with non-parametric and parametric methods [14].

In each frequency bin, the joint distribution of DFT coefficients of the two dependent zero-mean signals is modeled by the complex GMM as below:

$$p(\tilde{S}_{1,k,2,k}) = \sum_{i=1}^{Q} \omega_{i,k}\, p_G(\tilde{S}_{1,k,2,k}, \boldsymbol{C}_{i,k})$$
$$\boldsymbol{C}_{i,k} = \begin{bmatrix} \sigma^2_{1,i,k} & \sigma_{12,i,k} \\ \sigma^*_{12,i,k} & \sigma^2_{2,i,k} \end{bmatrix}, \sum_{i=1}^{Q} \omega_{i,k} = 1 \qquad (8)$$
$$k = 1, \cdots, L/2 - 1$$

where Q is the number of components in GMM, i is the index to the ith Gaussian components, $p_G(\tilde{S}_{1,k,2,k}, \boldsymbol{C}_{i,k})$ is the centered complex Gaussian distribution with covariance matrix $\boldsymbol{C}_{i,k}$, $\omega_{i,k}$ is the weight of the ith Gaussian component in the kth frequency bin.

By fixing the mean of each component as zero, the number of parameters to be estimated in each frequency bin is further reduced to $4Q$, including $3Q$ in the covariance matrix $\boldsymbol{C}_i(k)$ and Q in the weights ω_i.

The $4Q$ parameters are estimated in the first stage of the algorithm from samples of individual sources by standard expectation-maximization (EM) algorithm with K-means initialization. On-line EM algorithm can also be applied to enable a real-time implementation [15].

IV. SEPARATION OF MIXTURE

In the separation stage, the traditional Wiener filter [16] is extended in two aspects to separate the mixture of the two dependent sources. Firstly, it is generalized to take the correlation between the dependent sources into consideration. Secondly, the fixed gain of the Wiener filter is extended to be adaptive so that the local stationarity can be dealt with. These two aspects lead to the design of an adaptive weighted Wiener filter.

Based on the information obtained in the first stage, the two sources can be estimated through the MAP estimator in each frequency bin:

$$(\hat{\tilde{S}}_{1,k}, \hat{\tilde{S}}_{2,k}) = \arg\max_{\tilde{S}_{1,k}, \tilde{S}_{2,k}} p(\tilde{S}_{1,k}, \tilde{S}_{2,k} \,|\, \tilde{X}_k) \qquad (9)$$

However, (9) is not directly tractable [7]. To get back to the traditional Wiener filtering case, a hidden random variable q is introduced, which is associated with the active Gaussian component in GMM and is referred to as state variable. The posterior probability in (9) is then formulated as:

$$p(\tilde{S}_{1,k}, \tilde{S}_{2,k} \,|\, \tilde{X}_k) = \sum_{j=1}^{Q} p(\tilde{S}_{1,k}, \tilde{S}_{2,k} \,|\, \tilde{X}_k, q = j) p(q = j \,|\, \tilde{X}_k) \qquad (10)$$

Therefore, the estimation of sources needs three steps: estimate the current state by calculating the posterior probability of the state variable; construct the filters by maximizing the posterior probability of the sources given the state; separate the mixture and reconstruct the two sources in the time domain. In the following formula, $q = j$ is abbreviated as q_j.

A. State Estimation

The state variable q can be estimated through the posterior probability, i.e., $p(q_j \,|\, \tilde{X}_k)$, denoted as $\gamma_{j,k}$. It is calculated as:

$$\gamma_{j,k} \propto p(\tilde{X}_k \,|\, q_j) p(q_j) = p(\tilde{X}_k \,|\, q_j) \omega_{j,k} \qquad (11)$$

When the active state is given as j, $\tilde{X}_k$ is the sum of two correlated complex Gaussian variables with the joint distribution:

$$\left[\tilde{S}_{1,k}, \tilde{S}_{2,k}\right]^T \sim CN(\boldsymbol{0}, \boldsymbol{C}_{j,k}) \qquad (12)$$

where $\boldsymbol{C}_{j,k}$ is given in (8). It can be shown $\tilde{X}_k$ follows:

$$p(\tilde{X}_k \,|\, q_j) = p_G(\tilde{X}_k, \sigma^2_{1,j,k} + \sigma^2_{2,j,k} + 2\sigma_{12,j,k}) \qquad (13)$$

Inserting (13) into (11), we obtain,

$$\gamma_{j,k} \propto \omega_{j,k}\, p_G(\tilde{X}_k, \sigma^2_{1,j,k} + \sigma^2_{2,j,k} + 2\sigma_{12,j,k}) \qquad (14)$$

B. Construction of the filters

Given the active state q, (9) can be solved by extending the Wiener filter.

It is obvious that

$$p(\tilde{X}_k \,|\, \tilde{S}_{1,k}, \tilde{S}_{2,k}, q_j) = \delta(\tilde{S}_{1,k} + \tilde{S}_{2,k} - \tilde{X}_k) \qquad (15)$$

where $\delta(\bullet)$ is the Dirac delta function.

$p(\tilde{S}_{1,k}, \tilde{S}_{2,k} \,|\, q_j)$, the likelihood of the hidden q process can be calculated straightforward:

$$p(\tilde{S}_{1,k}, \tilde{S}_{2,k} \,|\, q_j) = p_G(\tilde{S}_{1,k,2,k}, \boldsymbol{C}_{j,k}) \qquad (16)$$

Therefore, given the active component in GMM, the posterior probability of the two sources is:

$$p(\tilde{S}_{1,k}, \tilde{S}_{2,k} \,|\, \tilde{X}_k, q_j) \propto p(\tilde{X}_k \,|\, \tilde{S}_{1,k}, \tilde{S}_{2,k}, q_j) p(\tilde{S}_{1,k}, \tilde{S}_{2,k} \,|\, q_j)$$
$$= \delta(\tilde{S}_{1,k} + \tilde{S}_{2,k} - \tilde{X}_k) p_G(\tilde{S}_{1,k,2,k}, \boldsymbol{C}_{j,k}) \qquad (17)$$

The MAP estimator (9)-(10) can be solved by picking up the Gaussian component with the highest probability calculated in (14) and maximizing (17) under the constraint:

$$\tilde{S}_{1,k} + \tilde{S}_{2,k} = \tilde{X}_k \qquad (18)$$

The solution can be easily found as:

$$\hat{\tilde{S}}_{1,k} = \frac{\sigma_{1,j,k}^2 + \mathrm{Re}(\sigma_{12,j,k})}{\sigma_{1,j,k}^2 + \sigma_{2,j,k}^2 + 2\,\mathrm{Re}(\sigma_{12,j,k})}\tilde{X}_k$$
$$\hat{\tilde{S}}_{2,k} = \frac{\sigma_{2,j,k}^2 + \mathrm{Re}(\sigma_{12,j,k})}{\sigma_{1,j,k}^2 + \sigma_{2,j,k}^2 + 2\,\mathrm{Re}(\sigma_{12,j,k})}\tilde{X}_k \quad (19)$$

where j corresponds to the component with highest $\gamma_{j,k}$ in the kth frequency bin.

An alternative estimator PM in (3) assigns every Gaussian component with a probability instead of a hard decision on active components, which is shown below:

$$\begin{aligned} E[\tilde{S}_{1,k}\,|\,\tilde{X}_k] &= \int_{\tilde{S}_{1,k}} (\tilde{S}_{1,k} \int_{\tilde{S}_{2,k}} p(\tilde{S}_{1,k}, \tilde{S}_{2,k}\,|\,\tilde{X}_k) d\tilde{S}_{2,k}) d\tilde{S}_{1,k} = \\ &\int_{\tilde{S}_{1,k}} (\tilde{S}_{1,k} \int_{\tilde{S}_{2,k}} (\sum_{j=1}^{Q} p(\tilde{S}_{1,k}, \tilde{S}_{2,k}\,|\,\tilde{X}_k, q_j) p(q_j\,|\,\tilde{X}_k)) d\tilde{S}_{2,k}) d\tilde{S}_{1,k} \\ &= \sum_{j=1}^{Q} \gamma_{j,k} (\int_{\tilde{S}_{1,k}} (\tilde{S}_{1,k} \int_{\tilde{S}_{2,k}} p(\tilde{S}_{1,k}, \tilde{S}_{2,k}\,|\,\tilde{X}_k, q_j) d\tilde{S}_{2,k}) d\tilde{S}_{1,k}) \\ &= \sum_{j=1}^{Q} \gamma_{j,k} E\left[\tilde{S}_{1,k}\,|\,\tilde{X}_k, q_j\right] \end{aligned} \quad (20)$$

Since $\tilde{S}_{1,k}$ follows a Gaussian distribution for a given active state q, its mean corresponds to the peak location. In other words, the MAP estimation (19) can replace $E\left[\tilde{S}_{1,k}\,|\,\tilde{X}_k, q_j\right]$ in (20). Therefore the PM estimation for the two sources is:

$$\hat{\tilde{S}}_{1,k} = \sum_{j=1}^{Q} \gamma_{j,k} \frac{\sigma_{1,j,k}^2 + \mathrm{Re}(\sigma_{12,j,k})}{\sigma_{1,j,k}^2 + \sigma_{2,j,k}^2 + 2\,\mathrm{Re}(\sigma_{12,j,k})}\tilde{X}_k$$
$$\hat{\tilde{S}}_{2,k} = \sum_{j=1}^{Q} \gamma_{j,k} \frac{\sigma_{2,j,k}^2 + \mathrm{Re}(\sigma_{12,j,k})}{\sigma_{1,j,k}^2 + \sigma_{2,j,k}^2 + 2\,\mathrm{Re}(\sigma_{12,j,k})}\tilde{X}_k \quad (21)$$

(21) can be regarded as a generalized Wiener filter. It separates the mixture by considering the correlation term between the signals. Besides, it is weighted by the posterior probability $\gamma_{j,k}$. Since $\gamma_{j,k}$ is adaptive for locally stationary signals, (21) is essentially a weighted adaptive Wiener filter.

V. SIMULATION RESULTS

The proposed algorithm is evaluated on a male speech s_1, which is filtered by a 128-tap impulse response shown in Figure 1 to form a linearly correlated source s_2. The impulse response has a shape of typical response of feedback path. It is chosen so that the correlation between s_1 and s_2 varies with frequencies, and the variances of s_1 and s_2 are comparable. The first 45 seconds of the two signals are used to train the GMM. The following 15 seconds are mixed for separation.

The PM estimator in (21) is selected for the simulation since in GMM it is usually superior to the MAP estimator as shown in [7]. The frame length is 512 samples, corresponding to approximately 5 milliseconds. The number of Gaussian components is 3.

The popular measures of source separation performance, such as in [17], usually apply for independent sources. For dependent sources, the measure adopted in this paper is a simple normalized test error, defined in the time domain as:

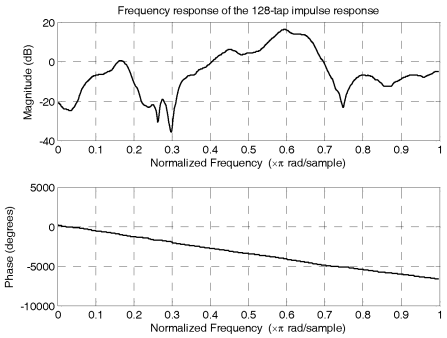


Fig. 1. Frequency response of the 128-tap echo-like impulse response

$$\varsigma_i = \frac{\left\|\hat{s}_i - s_i\right\|_2}{\left\|s_i\right\|_2}, i = 1, 2 \quad (22)$$

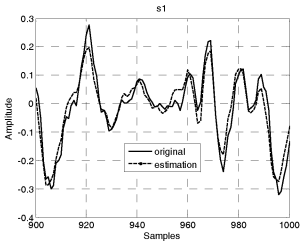
where $\|\cdot\|_2$ denotes the L2-norm.

The separation results are shown in Figure 2. Figure 2(a) and 2(b) illustrate the partial waveforms of the two sources. Figure 2(c) and 2(d) are original spectrograms of the two signals. Figure 2(e) and 2(f) are the estimated spectrograms. The original speech signal is cut off at 4 kHz. Although the estimated signal is not sharply cut off there due to a higher sampling rate (11025 Hz), the estimated frequency contents are small enough above 4 kHz. The comparison between the performance of traditional Wiener filter and the proposed algorithm is given in Table 1. Figure 2 and Table 1 show that the proposed algorithm can separate the excerpted signals very well.

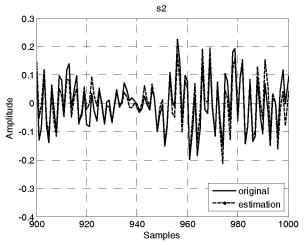
It is also noted that there exists a pattern in Figure 2(e) and 2(f): Horizontal broken lines are located at some evenly spaced frequency bins. This indicates the failure of modeling at these frequencies. One possible reason is that voiced speech shows strong tonal characteristics at harmonic frequencies. The DFT coefficients at these frequencies tend to be constant, which GMM with zero-mean components is not able to model. The other possible reason is that the proposed algorithm is phase blind, which is inherited from Wiener filtering. In some frequency bins, the two sources could be negatively correlated. The mixture is therefore the remaining signal after mutual cancellation. The amount of cancellation is impossible to recover when the phase information is missing. The separation performance is thus severely degraded in those frequency bins.

Table 1. Comparison between the performance of Wiener filter and generalized Wiener filter (three Gaussian components)

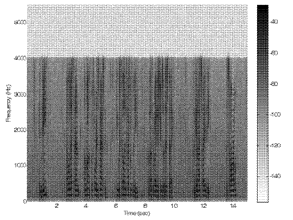
Normalized Error	Source 1	Source 2
Wiener Filter	0.4060	0.4351
Generalized Wiener Filter	0.3497	0.3742



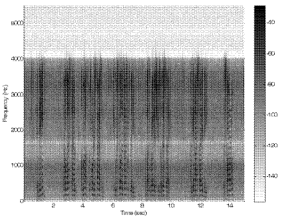
(a) partial waveform of source 1



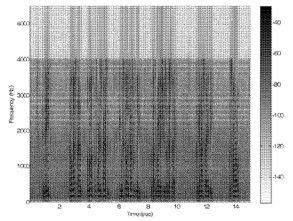
(b) partial waveform of source 2



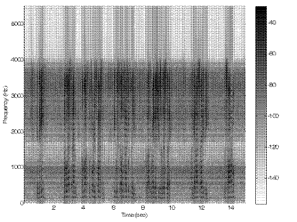
(c) original spectrogram of source 1



(d) original spectrogram of source 2



(e) estimated spectrogram of source1



(f) estimated spectrogram of source2

Fig. 2. Simulation results with 3 Gaussian components and 512-sample frame length ($\varsigma_1 = 0.3497, \varsigma_2 = 0.3742$)

The separation performance as a function of the frame size and the number of Gaussian components is studied. In Figure 3, as the number of Gaussian components increases, the normalized test error drops down first and increases later. The drop indicates the inadequacy of Gaussian components in describing the two locally stationary dependent signals. The following rise is probably a consequence of over-fitting. The optimal number for the excerpted signals is three. In Figure 4, it is seen that the performance is generally better with a finer spectrum, i.e., longer frames. Since time domain modeling can be regarded as a special frequency-domain modeling when the frame length is equal to 1, this infers that time-domain modeling is worse than frequency-domain modeling.

It can be noted that the test error of s_2 is consistently higher than s_1 . This is because s_1 has a larger variance than s_2 . Constraint (18) requires:

$$x(n) = s_1(n) + s_2(n) = \hat{s}_1(n) + \hat{s}_2(n)$$
$$s_1(n) - \hat{s}_1(n) = s_2(n) - \hat{s}_2(n) \quad (23)$$

Therefore,

$$\frac{\|s_1 - \hat{s}_1\|_2}{\|s_1\|_2} < \frac{\|s_2 - \hat{s}_2\|_2}{\|s_2\|_2}, if \ \|s_1\|_2 > \|s_2\|_2 \quad (24)$$

i.e.,

$$\varsigma_1 < \varsigma_2 \quad (25)$$

Thus the dominant source is always estimated better. The more dominant the better estimated.

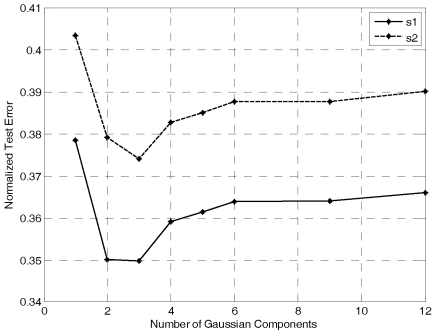


Fig. 3. Normalized test error as a function of the number of Gaussian components

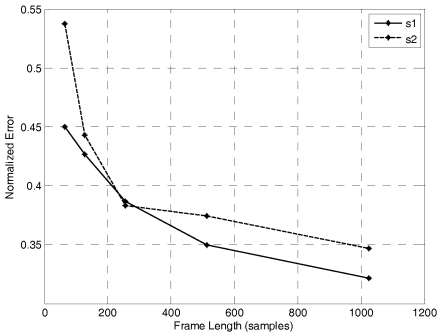


Fig. 4. Normalized test error as a function of frame length

VI. CONCLUSIONS AND FUTURE WORK

A two-stage algorithm for monaural separation of dependent audio sources is proposed in this paper. The algorithm is based on the Bayesian framework and designed for general audio sources, not restricted to speech or music. In the first stage of the algorithm, complex GMM is used in the frequency domain to capture the properties of the signals and their correlation. In the second stage, based on the information obtained, traditional Wiener filter is extended to an adaptive weighted Wiener filter to separate the mixture. The extension takes both correlation term and local stationarity into account. The simulation results indicate that the proposed algorithm performs very well when the two sources are linearly correlated and have comparable variances.

As explained in (23)-(25), when the two sources have incomparable variances, i.e. one of the sources dominates, the stronger source is always estimated better. This is an inherent property of the measure and constraint (18). It could also be regarded as one of the limitations in many source separation algorithms including the proposed algorithm in this paper.

Due to the loss of phase information in Wiener filtering, when the two sources are negatively correlated, it is very hard to estimate the amount of cancellation in the mixture for an under-determined source separation problem. An additional phase model could be added to the training step of the algorithm to obtain the phase information. A relatively easy way is to model the phase difference between the two signals in each frequency bin because it is usually more stable than the phases of individual sources, especially when the correlation of the two sources is time-invariant or slowly time-varying. A successful phase modeling should improve the performance of the algorithm significantly.

Another aspect of future work is to evaluate the algorithm with non-linearly correlated sources. This may lead to some modifications of the algorithm, such as modeling the correlation between adjoining frequency bins.

It should be noted that the proposed algorithm needs more signal samples to assess the performance before conclusions about the best choice of components and frame length are drawn.

REFERENCES

[1] L. Bedini, D. Herranz, E. Salerno, C. Baccigalupi, E. E. Kuruoglu, and A. Tonazzini, "Separation of Correlated Astrophysical Sources Using Multiple-Lag Data Covariance Matrices", *EURASIP Journal on Applied Signal Processing*, vol. 2005, Issue 15, pp. 2400-2412.

[2] Frederic Vrins, John A. Lee, and Michel Verleysen, "Filtering-Free Blind Separation of Correlated Images", *the 8th International Workshop on Artificial Neural Networks*, 2005.

[3] Cichocki A., and Georgiev P., "Blind Separation Algorithms with Matrix Constraints", *IEICE Transactions Fundamentals of Electronics Communications and Computer Science*, E86-A, pp. 522-531, 2003.

[4] F. Abrard, and Y. Deville, "A Time-frequency Blind Signal Separation Method Applicable to Underdetermined Mixtures of Dependent Sources", *Signal Processing*, vol. 85, pp. 1389-1403, July 2005.

[5] F. Abrard, Y. Deville, "Blind Separation of Dependent Sources Using the Time-frequency Ratio of Mixtures Approach", *the 7th International symposia Signal Processing Applications (ISSPA)*, July 2003.

[6] Yuanqing Li, Shun-Ichi Amari, Andrzej Cichocki, Daniel, W. C. Ho, and Shengli Xie, "Underdetermined Blind Source Separation Based on Sparse Representation", *IEEE Transactions on Signal Processing*, vol.54, no. 2, pp. 423-437, 2006.

[7] Benaroya, L., Bimbot, F., and Gribonval, R., "Audio Source Separation with a Single Sensor", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 191-199, January 2006.

[8] Daniel P.W.Ellis, Ron J.Weiss, "Model-based Monaural Separation Using a Vector-Quantized Phase-Vocoder Representation", *ICASSP 2006*.

[9] Thomas Melia and Scott Rickard, "Underdetermined Blind Source Separation in Echoic Environments Using DESPRIT", *EURASIP Journal on advances in Signal Processing*, vol. 2007, Article ID 86484.

[10] Steven M. Kay, "Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory", *Prentice Hall PTR*, 1993.

[11] William A. Pearlman, and Robert M. Gray, "Source Coding of the Discrete Fourier Transform", *IEEE Transactions on Information Theory*, vol. 24, no. 6, pp. 683-692, November 1978.

[12] R. A. Wooding, "The Multivariate Distribution of Complex Normal Variables", *Biometrika*, vol. 43, no. 1-2, pp. 212-215, 1956.

[13] Laurent Benaroya, and Frederic Bimbot, "Wiener Based Separation with HMM/GMM Using a Single Sensor", *Proceedings 4th International Symposia on Independent Component Analysis and Blind Signal Separation*, April 2003.

[14] Christopher M. Bishop, "Neural Networks for Pattern Recognition", pp. 33, *Oxford University Press*, November 1995.

[15] Masa-aki Sato, and Shin Ishii, "On-line EM Algorithms for the Normalized Gaussian Network", *Neural Computation*, vol.12, no. 2, pp. 407-432, 2000.

[16] Wiener, Norbert, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series", *The MIT Press*, 1949.

[17] E. Vincent, R. Gribonval, and C. FEVOTTE, "Performance Measurement in Blind Audio Source Separation", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462-1469, July 2006.